中文分词模型词典融入方法比较 *

冯　雪

(北京信息科技大学 计算机学院，北京 100192)

摘　要：目前比较流行的中文分词方法为基于统计模型的机器学习方法。基于统计的方法一般采用人工标注的句子级的标注语料进行训练，但是这种方法往往忽略了已有的经过多年积累的人工标注的词典信息。这些信息尤其是在面向跨领域时，由于目标领域句子级别的标注资源稀少，从而显得更加珍贵。因此如何充分而且有效的在基于统计的模型中利用词典信息，是一个非常值得关注的工作。最近已有部分工作对它进行了研究，按照词典信息融入方式大致可以分为两类：一类是在基于字的序列标注模型中融入词典特征，而另一类是在基于词的柱搜索模型中融入特征。对这两类方法进行比较，并进一步进行结合。实验表明，这两类方法结合之后，词典信息可以得到更充分的利用，最终无论是在同领域测试和还是在跨领域测试上都取得了更优的性能。

关键词：中文分词；条件随机场；柱搜索；领域自适应

中图分类号：TP391.1　　doi: 10.3969/j.issn.1001-3695.2017.05.0643

Comparison of methods for integrating lexicon information in Chinese word segmentation

Feng Xue

(*School of Computer*, *Beijing Information Science* & *Technology University*, *Beijing* 100192, *China*)

Abstract: Chinese word segmentation is a fundamental task in Chinese natural language processing. Currently the mainstream methods for Chinese word segmentation exploit statistical machine learning models. These methods usually require manual-annotated segmented sentences as training corpus, yet have neglected the annotated large-scale lexicon resources which have been built before, where these resources can be highly valuable when cross-domain evaluation is conducted, as the gold-standard sentence-level annotations arerare. Recently, the integration of lexicon formation into word segmentation models has gained increasing interest. As a whole, the integration methods can be classified into two categories: one being based on character-based models that cast word segmentation problem as sequence labeling, and the other being based on word-based models that use beam-search to decode. In this paper, we compare these two models, and combine them. Experimental results on benchmark data sets show that lexicon information can be more fully explored after combination, and finally the combined model can achieve better performances with both in- and cross-domain settings.

Key Words: Chinese word segmentation; conditional random field; beam-search; domain adaption

0　引言

中文和其他字母形式的语言的一个巨大区别是中文输入的词与词之间并不存在显式的词语分隔符，因此在进行中文自然语言处理时，一般最首要的任务是进行中文分词。当分词完毕之后，中文自然语言处理的各项分析以及应用便可以像其他语言一样，用一套统一相似的框架展开。中文分词的研究工作已经持续了很长的时间，目前比较流行而且性能比较好的方法是基于统计机器学习的方法。

中文分词的统计模型中，主流的方法分为两种，一种是基于字序列分类的方法[1-3]，而另一类为基于词的搜索算法[4-5]。基于字序列分类的方法将中文分词转换成序列标注任务：将句子中的每一个字用 BMES 四个类型的标签表示，其中 B 表示一个词的首字，M 表示一个词的中间字，E 表示一个词的结尾字，而 S 表示该词是由一个单一的字组成。通过这样的转换之后，便可以采用标注的条件随机场模型进行训练和解码，从而完成中文分词任务。

由于在基于字的序列标注的方法中，与词相关的特征难以融入，所以后续有人提出了基于词的方法，这一方法在解码时，对下一个字符的处理采取两个动作，要么和前面的词进行拼接合并，要么自己成为下一个词的首字，这样解码时，模型便能知道过去产生了哪些具体的词，从而这些词也可以作为模型的

基金项目：北京市教委科技计划面上项目（KM201411232012）

作者简介：冯雪（1984-），女，江苏徐州人，讲师，博士，主要研究方向为数字版权保护技术、自然语言处理（fengxue@bistu.edu.cn）.

特征。由于每个字都面临两个选择，这样搜索空间随着处理字符数目的增加便会以指数级别增长，因此这种方法往往采用柱搜索的算法来解码。

上面两种典型的方法都是利用句子级别的分词信息进行模型训练，往往忽略了过去长期积累的词典资源，而且另一方面词典的标注实际上要比句子级分词的标注要容易很多。如何在基于统计的中文分词模型中充分的利用词典资源也是非常重要的研究内容。显然，上面两种典型的方法，由于模型上的显著差异，它们融入词典的方式肯定是不一样的。其中，张梅山等人在 2011 年提出了一种针对序列标注模型的词典融入方法[6]；而基于词的模型方法中，词典融入显得更容易一些，直接判断最近形成的一个词是否在词典中即可，这一方法也被 Zhang 等人于 2014 年使用在一个分词词性标注的联合模型中[7]。

本文的研究目的在于系统的比较这两种融入词典信息的方法，探索它们的差异性，并进一步将这两种方法进行结合，观察这一结合能否带来更好的效果。最后，本文通过实验对这两种方法进行了对比，采用了两种设置，即同一领域和跨领域。结果发现，基于词典的方法不仅词典融入方式更简单，而且还能带来更好的效果；当两种方法结合之后，无论什么设置，都能得到最好的效果。

1 相关工作

中文分词在过去已经得到研究者大量的研究[1-5,8-10]。最开始的工作主要采用基于规则的方式，以及采用语言模型的方法对分词结果进行打分。最近几年，采用统计机器学习的方法逐渐取得了领先的性能，这一方法主要包括两类：基于字的序列标注的方法和基于词的方法。这两类方法各有利弊，孙薇薇在 2010 年 COLING 上对这两种方法进行了详细的比较和分析，然后进一步将这两种方法进行了结合[11]，这一比较融合的思想和本文比较相似。但是和上述工作的主要区别在于，本文主要关注词典信息的充分利用，针对基于词典的特征提出了融合，进一步特别关注了词典特征在跨领域条件下的性能。

统计模型中融入词典信息最早是由赵海等人[12]提出的，进一步张梅山等人[6]对他们的方法进行扩展，使得这一方法能够在跨领域上发挥作用。他们的方法主要是在基于字的序列标注的方法上展开的。进一步 Zhang 等人[7]在基于词的方法上也尝试了融入词典特征。本文的思路和孙薇薇等人思路一致，但是主要针对的是对于词典信息的融入这一特性进行详细对比以及进一步的结合。词典信息的融入对于中文分词是非常重要的，尤其是在跨领域方面，训练语料不足的情况下，这一信息的合理融入能带来更好的性能。

2 基于字的序列标注模型

2.1 CRF 中文分词模型

将中文分词当作序列标注这一问题，最早是由薛念文等人在 2003 年提出的，后续有人对此方法进行了改进。目前，比较流行的设置是采用 BMES 四类标记的方法，将句子中每个字根据它在词中的位置进行分类：其中 B 表示一个词的开始字符，M 表示一个词中间位置的字符，E 表示一个词的结尾字符而 S 表示一个词由独立的一个字构成。

CRF 是条件随机场(Conditional Random Field)的简称，它是目前最主流的序列标注算法，因为这一方法在序列标注问题上能取得领先的效果。对于给定的一个句子 $x=c_1 \ldots c_n$ 及其一个分词结果为 $y=y_1 \ldots y_n$，其中 c_i 为中文字符，$y_i \in \{B,M,E,S\}(1 \le i \le n)$，则可以用如下的方法计算 y 的分数，这一分数实际上也是一个概率值：

$$P_W(y \mid x) = \frac{1}{Z(x)} \exp(W \cdot \sum_{i=1}^{n} \Phi(y_{i-1}, y_i, x))$$

其中：$Z(x)$ 是归一化因子，$\Phi(y_{i-1}, y_i, x)$ 为特征向量函数，W 为特征权重向量。

模型中使用的特征主要包括字的一元、二元和三元特征，以及字符类别特征，具体定义可以参考张梅山等人 2011 年的论文，这里不进行详细介绍。

2.2 词典信息的融入

统计机器学习的模型中，实际上最重要的部分是特征选择，因为这类方法的核心是学习特征在打分时的权重，因此在统计模型中融入词典信息实际上就转换为如何将词典相关的特征加入到模型中。这里本文直接介绍张梅山等人 2011 年提出的一系列特征。

在介绍具体特征之前，首先需要定义三类函数。对于给定句子 $x=c_1 \ldots c_n$，以及词典 D，考虑其中的第 j 个字符 $c_j (1 \le j \le n)$，定义如下三个函数：

$$f_B(x,j,D) = \max l\,,$$
$$s.t.\begin{cases} w = c_j \ldots c_{j+l-1} \in D \\ j+l-1 \le n \end{cases}$$

$$f_M(x,j,D) = \max l\,,$$
$$s.t.\begin{cases} w = c_s \ldots c_{s+l-1} \in D \\ j < s+l-1 \le n \\ 1 \le s < j \end{cases}$$

$$f_E(x,j,D) = \max l\,,$$
$$s.t.\begin{cases} w = c_{j-l+1} \Lambda\ c_j \in D \\ 1 \le j-l+1 \end{cases}$$

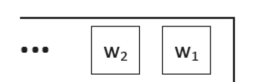
其中：w 表示一个词语；$f_B(x,j,D)$ 表示对于句子 x 在 j 位置根据词典 D 采用正向最大匹配所获得的词的长度；$f_M(x,j,D)$ 表示对于句子 x 在 j 前面的某个位置根据词典 D 采用正向最大匹配所获得的经过 j 位置而且不以 j 结尾的最长词的长度；$f_E(x,j,D)$ 表示对于句子 x 在 j 位置根据词典 D 采用逆向最大匹配所获得的词的长度。

定义了上述的三个函数之后，便可以利用这些值来定义词典特征。对于第 i 个位置，本文使用的特征主要包括 $f_B(x,k,D)$，$f_M(x,k,D)$，$f_E(x,k,D)$，$k=i, i \pm 1$，以及它们的各种组合，主要包含二元和三元的组合。

3 基于词的柱搜索模型

3.1 基本模型介绍

基于词的柱搜索算法最早由张岳和 Clark[4]于 2007 年提出来，进一步在张岳和 Clark(2011)[5]中得到完善。它又通常被称为基于转移的模型，这一模型中，其核心是将解码转换成一个动作序列，每执行一个动作，则解码的进程又称为状态会发生改变。具体，解码过程中的每个状态由一个栈和一个队列组成，栈中存储这一个已经部分解码的中文词序列，而队列中存储的是尚未进行处理字序列，如图 1 所示。动作分为两类，一类是 SEPARATE，表示将队列中的第一个字移入栈中，作为下一个词的开始；而另一类为 Append，即将队列中的第一个字附加到栈顶的那个词后。解码初始时，栈为空而队列中存储句子中的所有字，解码结束时，队列为空，栈中的结果即为最终的分词结果。



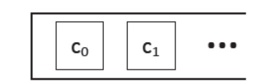


图 1 基于词的柱搜索模型中定义的状态示意图

上述算法在解码时，每个状态都能变为两个状态，因此在处理到第 i 个字符时，生成的可能的状态个数为 2^i，因此算法的时空复杂度是指数级别增长的，为了克服这一缺点，该方法进一步采用了柱搜索算法，即每次只保留分数最高的固定大小的状态数目。

具体分数的计算公式比较简单，直接将每个动作所产生的特征累加合并，然后将得到的稀疏向量和模型参数点乘即可：

$$score(x,a_1,\Lambda,a_n)=W\sum_{i=1}^{n}\Phi(a_i,x)$$

其中：W 为模型参数，$\Phi(a_i,x)$ 为特征抽取函数，具体特征的定义可以参考张岳和 Clark(2011)[5]。模型参数训练方法采用平均感知机算法，具体细节这里不再详细介绍。

3.2 词典信息的融入

在基于词的柱搜索模型中，对于词典信息相关特征的融入和基于字的方法比较起来，由于该模型能直接看到生成的词的信息，因此显得方便很多。本文仿效 Zhang 等人(2014)年的方法，在执行 SEPARATE 操作时，判断栈顶刚生成的词是否在词典 D 中出现，以及该词的长度为多少。

4 模型对比和结合

基于字的序列标注模型和基于词的柱搜索模型在融入词典特征信息方面，存在着很大的相似性。a)两个模型提取的特征都是和其中具体的词无关的特征，因此即便词典在测试阶段更换之后，仍然能够使用;b)匹配词的长度信息实际都显式的编码在模型中，这一直觉和传统最早的正向最大匹配算法的直觉非常一致;c)由于特征的去词汇化，这两个模型在领域切换方面可以非常灵活，图 2 显示了这两个模型在领域切换时的训练和解码模式：用户只需要训练一个模型，在跨领域测试时，只需在解码时将目标领域的词典换上即可。

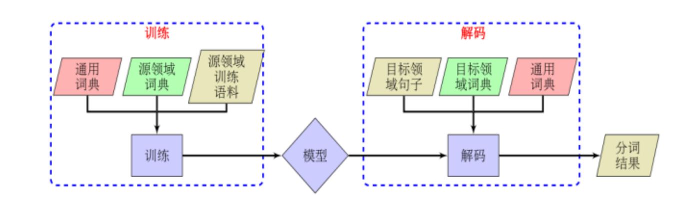


图 2 跨领域训练与解码示意图

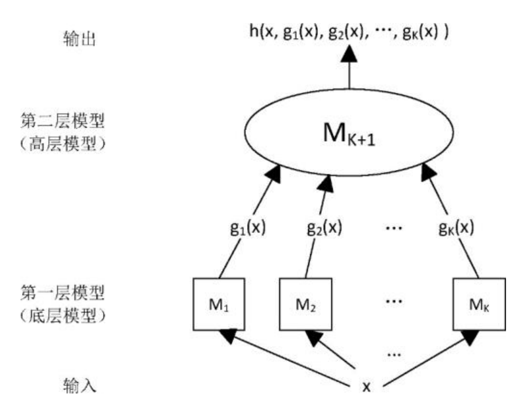


图 3 基于栈的模型融合框架示意图

基于字的序列标注模型和基于词的柱搜索模型的不同点主要体现在模型本身上，首先是具体特征不一样，因为基于字的模型无法直接的获取词，因此特征的定义是间接的反映词典中的词信息；而基于词的方法则可以直接对比词典中的词和目前刚产生的词是否一致。其次，特征训练方式不一样，基于字的方法采用的条件随机场，实质上时一个概率图模型；而基于词的方式采用平均感知机，属于最大间隔算法。

关于这两个模型的结合，这里直接采用基于栈的方式进行结合，即将其中给一个模型的结果传送给第二个模型，作为特征即可，如图 3 所示。这一方式已经广泛的应用于模型融合上面，由于它在理论上显得更优雅，而且也能带来更好的性能。具体实现上，这里将基于字的模型的结果传递给基于词的模型，虽然也可以将基于词的模型的结果传递给基于字的模型，但是本文通过初步实验发现，这种形式的结合能取得更好的性能。

5 实验

本文使用和张梅山等人(2011)相同的语料进行模型训练和测试，利用 SIGHAN BAKEOFF 2005 中的 PKU 训练语料进行训练，测试时包含两个领域，其中一个为 PKU 领域，而另一个数据来自于金融领域，主要是为了测试模型在本领域和跨领域时的性能。通用词典来自于北京大学中国语言学研究中心公开的词典[1]，一共包含大约 10 万多个词，而 PKU 领域词典和金融

[1]http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/Source_Code/Chapter_8/Lexicon_full_2000.zip

领域词典则分别从训练语料中构建和金融相关的百度百科中进行抽取和标注。

本文使用准确率(P)、召回率(R)和 F-measure 值(F)三个指标来评价分词模型，其中最重要的是 F-measure 值。

5.1　相同领域性能

首先，本文观察同一领域融入词典信息后的性能，并和不使用词典信息的模型进行对比，最终结果如表 1 所示。

表 1　PKU 领域分词性能对比。

模型	是否使用词典	P	R	F
基于字的模型	-词典	94.3%	95.4%	94.8%
	+词典	96.6%	96.6%	96.6%
基于词的模型	-词典	94.7%	95.1%	94.9%
	+词典	97.2%	96.9%	97.0%
两者结合	-词典	95.2%	95.6%	95.4%
	+词典	97.6%	97.5%	97.5%

从表 1 中结果可以看出，基于词的模型和基于字的模型在不使用词典信息时，两者性能相当；但是在使用词典信息后，基于词的模型性能要比基于字的模型高出 0.4%。当基于词的模型和基于字的模型结合后，无论是在使用词典还是在不使用词典的情况下，都取得了最好的性能。另外，使用词典之后，三种模型的分词性能都比不使用词典显著增强，这也表明了词典信息的有用性。

5.2　跨领域性能

前面的实验比较了基于字的模型和基于词的模型在同领域时使用词典和不使用词典的性能，同时给出了这两种模型进行结合后的性能。这里，本文进一步观察在跨领域的情况下，最终的结果是否和同领域的趋势完全吻合。

表 2 给出了最终的实验结果。通过实验结果的对比可以发现，跨领域情况下的确实和同领域的趋势完全吻合。值得注意的是，这一部分实验测试时，并没有重新训练模型，最大的改动就是原有 PKU 领域的词典被替换成了金融领域的词典。另外，从实验结果中看出，通过融入外部词典信息，跨领域的性能可以提升接近 7%，从原有的 87%左右提升到了 94%以上。

表 2　金融领域分词性能对比

模型	是否使用词典	P	R	F
基于字的模型	-词典	84.0%	89.7%	86.8%
	+词典	93.2%	93.5%	93.3%
基于词的模型	-词典	84.8%	89.2%	87.0%
	+词典	94.2%	93.4%	93.8%
两者结合	-词典	85.2%	90.0%	87.6%
	+词典	94.9%	94.0%	94.4%

5.3　模型对比分析

前面从实验结果中可以发现模型融合能来更好的效果，这里进一步通过对比这两个模型的错误分布，来观察这两个模型的不同。

图 4 给出了融入了词典信息之后，基于字的模型和基于词的模型在同领域设置和跨领域设置下的错误分布图。图 4 中的错误分布式根据单个句子的性能来计算的，其中横坐标给出了基于词的模型的性能，而纵坐标表示基于字的模型的性能，图中的每个散点代表一个句子。观察图 4，可以发现两个子图中的散点分布非常分散，都不在一条直线上，这表明了两则的错误分布对比是杂乱无章的，从而表明了这两个模型的差异性。

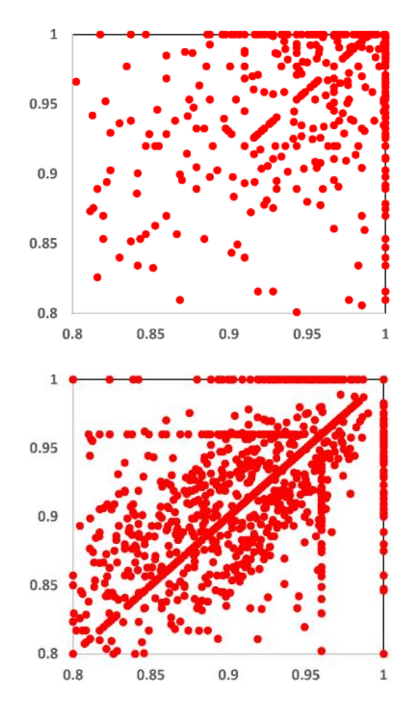


图 4　基于字的模型和基于词的模型错误分布图

6　结束语

本文对两类主流的分词模型在融入词典信息的方式上进行了介绍，并进行了对比和结合。本文不仅在模型建立的角度对两者进行了详细分析和比较，指出了两种模型在利用词典信息上的相同点和不同点，而且也从实验上对两个模型的错误分布进行了分析。分析结果表明两类模型虽然在词典融入方式上比较相近，但是也存在一定的差异性，因此这也表明模型融合会带来进一步的性能提升。本文通过基于栈的融合方式将两个模型进行了结合，而且最终实验结果表明两中方法的结合能够更有效的提升模型的最终性能。

本文的观点进一步验证了词典的信息对领域自适应是非常有效的，但是实际上词典的获取还是存在一定难度的， 虽然过去已经有了相当的积累。下一步工作集中在如何自动的获取某一领域的高质量词典。

参考文献：

[1] Xue Nianwen. Chinese word segmentation as character tagging [J]. International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8 (1): 29-48.

[2] Tseng H, Chang Pichuan, Andrew G, et al. A conditional random field word segmenter for SIGHAN bakeoff [C]// Proc of the 4th SIGHAN Workshop on Chinese Language Processing. 2005: 168–171.

[3] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th ICML. San Francisco: Morgan Kaufmann Publishers, 2001: 282–289.

[4] Zhang Yue, Clark S. Chinese segmentation with a word-based perceptron algorithm [C]// Proc of the 45th ACL. 2007: 840–847.

[5] Zhang Yue, Clark S. Syntactic processing using the generalized perceptron and beam search [J]. Computational linguistics, 2011, 37 (1): 105-151.

[6] 张梅山，邓知龙，车万翔，等. 统计与词典相结合的领域自适应中文分词 [J]. 中文信息学报, 2012, 26 (2): 8－12.

[7] Zhang Meishan, Zhang Yue, Che Wanxiang, et al. Type-supervised domain adaptation for joint segmentation and POS-tagging [C]// Proc of EACL. 2014: 588-597.

[8] Chang Pichuan, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance [C]// Proc of the 3rd Workshop on Statistical Machine Translation. Stroudsburg: Association for Computational Linguistics, 2008: 224-232.

[9] 许华婷，张玉洁，杨晓晖，等. 基于 active learning 的中文分词领域自适应 [J]. 中文信息学报, 2015, 29 (5): 55-63.

[10] 刘一佳，车万翔，刘 挺，等. 基于序列标注的中文分词、词性标注模型比较分析 [J]. 中文信息学报, 2013, 27 (4): 30-37.

[11] Weiwei Sun. Word-based and character-based word segmentation models: Comparison and combination [C]// Proc of the 23rd International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 1211-1219.

[12] Zhao Hai, Huang Changning, Li Mu. An improved Chinese word segmentation system with conditional random field [C]// Proc of the 5th SIGHAN Workshop on Chinese Language Processing. 2006: 162–165.